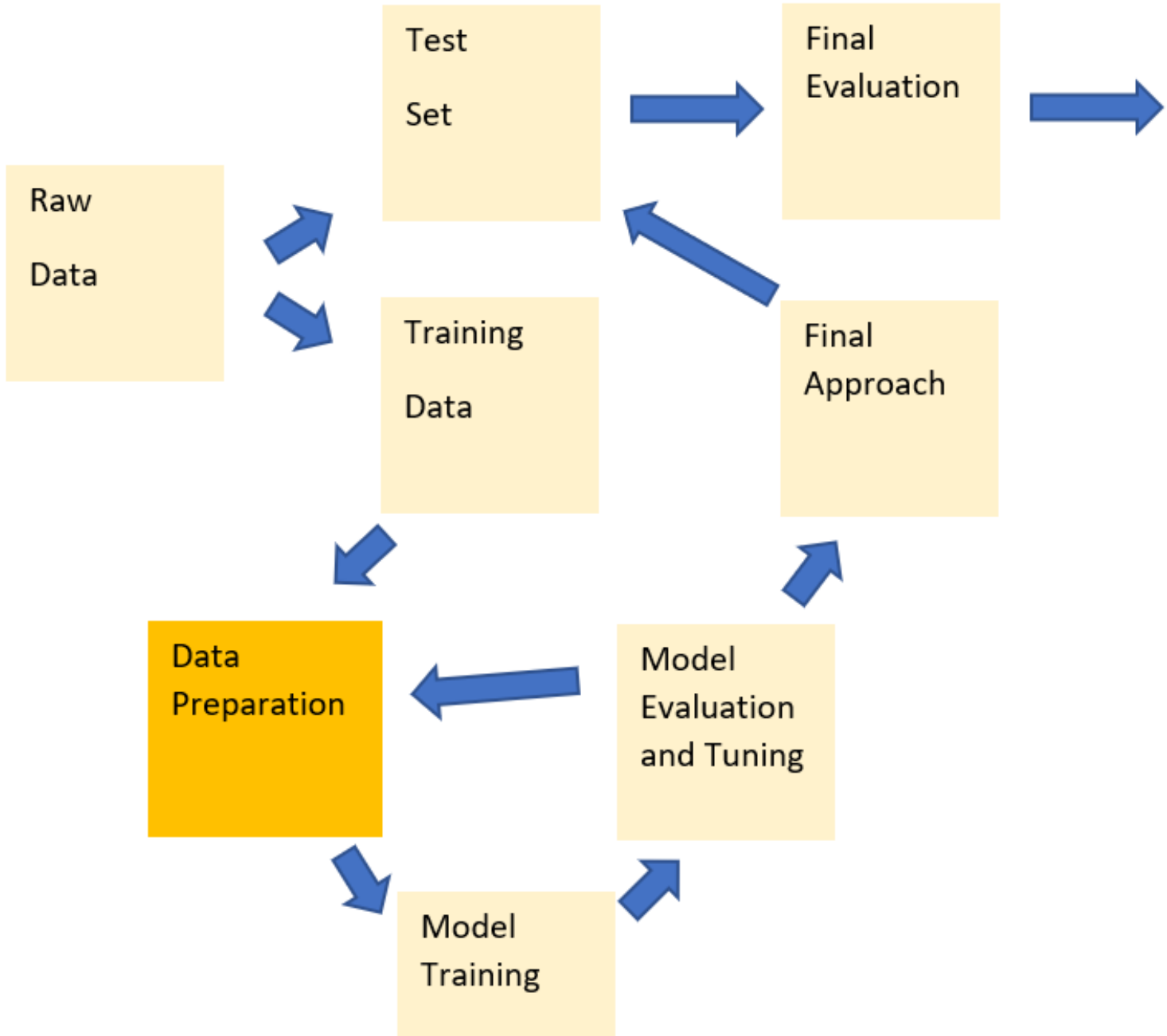
Additional Data Preparation Steps

Ryan Miller

Introduction



Data Preparation

Data preparation can be the most time-consuming aspect of any machine learning project, but also provides one of the best opportunities for performance improvement. Preparation steps typically include:

1. Feature engineering (today's lab)
2. Feature selection (future topic)
3. Imputing or excluding missing data (today's lecture)
4. Standardizing, scaling, and/or transforming features (previously covered)

Types of Missing Data

1. Missing Completely at Random (MCAR) - missing values are completely independent of the data itself
   - When data are MCAR, observations with any missing values can be discarded, or imputation can used.

Types of Missing Data

1. Missing Completely at Random (MCAR) - missing values are completely independent of the data itself
   - When data are MCAR, observations with any missing values can be discarded, or imputation can used.
2. Missing at Random (MAR) - missing values are not purely randomly, but they can be accounted for using available data
   - Under MAR, *imputation* should be used to retain the entire sample.

Types of Missing Data

1. Missing Completely at Random (MCAR) - missing values are completely independent of the data itself
   - When data are MCAR, observations with any missing values can be discarded, or imputation can used.
2. Missing at Random (MAR) - missing values are not purely randomly, but they can be accounted for using available data
   - Under MAR, *imputation* should be used to retain the entire sample.
3. Missing not at Random (MNAR) - the presence/absence of a missing value depends upon the missing value itself
   - Under MNAR, any modeling should be done with caution.

Examples of Missing Data

- MCAR: some blood samples are damaged by the handling processes used in a lab
- MAR: young people are less likely to have blood data available *because of their age* (which is recorded)
- MNAR: some blood samples are missing *because of the characteristics of those samples* (ie: high blood sugar is the reason for missing blood sugar data)

Imputation

Imputation is the processing of replacing a missing value with a substituted value. We'll briefly discuss two imputation strategies:

1. Simple imputation - each missing value is replaced by a constant, such as the mean/median/mode of that variable
2. Nearest neighbors imputation - each missing value is replaced using the k nearest data-points with available data

Simple Imputation

Pros:

- Computationally simple and convenient
- Easy to understand and implement

Cons:

- Introduces bias by changing the distributions of variables with missing data
- Especially problematic for nominal categorical variables

Nearest Neighbors Imputation

Pros:

- Less prone to introducing bias when compared with simple imputation

Cons:

- More computationally burdensome, and doesn't scale well to high-dimensional data
- Involves more subjective decisions (ie: number of neighbors, distance or uniform weighting, etc.)

Recommendations

- If the number of observations is large, and the amount of missing data is modest, excluding incomplete observations or using imputation are both reasonable strategies.
- If the number of observations is small, or if the amount of missing data is high, imputation should be used.
- If you suspect your data are MNAR, you should approach with caution (ie: consider the degree and direction of possible bias, the possibility of acquiring new data, etc.)